Statistical Inference

General linear regression model: \( Y_i \) is multivariate normal, with

\[
E(Y_i) = X_i \beta
\]

\[
\text{Cov}(Y_i) = \Sigma_i = \Sigma_i(\theta),
\]

where \( \theta \) is a \( q \times 1 \) vector of covariance parameters:

**unstructured:** the \( n(n + 1)/2 \) elements of the common \( \Sigma \);

**compound symmetry:** the two variance parameters \( \sigma_b^2 \) and \( \sigma_e^2 \).
The Likelihood

• Independence: for a single response $y_{i,j}$:

$$f_{i,j}(y_{i,j}) = (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (y_{i,j} - X'_{i,j}\beta)^2 \right\}$$

• Jointly, in logarithms:

$$l = \log \left\{ \prod_{i=1}^{N} \prod_{j=1}^{n} f_{i,j}(y_{i,j}) \right\}$$

$$= -\frac{Nn}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{i,j} - X'_{i,j}\beta)^2$$
Independence, continued:

• To maximize $l$ with respect to $\beta$, must minimize the sum of squares $\sum \sum (y_{i,j} - X'_{i,j}\beta)^2$.

• Least squares solution:

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n} (X_{i,j}X'_{i,j}) \right\}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{n} (X_{i,j}y_{i,j})$$
Correlated Observations

• Log likelihood:

\[ l = -\frac{1}{2} \sum_{i=1}^{N} \log(2\pi \Sigma_i) - \frac{1}{2} \sum_{i=1}^{N} (y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta). \]

• To maximize \( l \) with respect to \( \beta \), must minimize the “sum of squares”

\[ \sum_{i=1}^{N} (y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta). \]
Correlated Observations, known $\Sigma_i$ (GLS estimator):

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} \left( X'_i \Sigma_i^{-1} X_i \right) \right\}^{-1} \sum_{i=1}^{N} \left( X'_i \Sigma_i^{-1} y_i \right)$$

- unbiased:

$$E(\hat{\beta}) = \beta$$

- covariance:

$$\text{Cov}(\hat{\beta}) = \left\{ \sum_{i=1}^{N} \left( X'_i \Sigma_i^{-1} X_i \right) \right\}^{-1}$$
Correlated Observations, unknown $\Sigma_i$ (or $\theta$):

- Get estimate $\hat{\theta}$, e.g. by maximum likelihood or restricted maximum likelihood (REML);

- Plug in:

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} (X_i' \hat{\Sigma}_i^{-1} X_i) \right\}^{-1} \sum_{i=1}^{N} (X_i' \hat{\Sigma}_i^{-1} y_i)$$

- In large samples, approximately the same mean and covariance as if $\Sigma_i$ known.
Missing Data

Often parts of the data are missing. We model this by assuming that all responses were potentially observable, but some random mechanism decided which we actually observe. Two types of random mechanism must be distinguished.

**MCAR:** Data are *Missing Completely at Random* if missingness is independent of all parts of the complete set of responses.

**MAR:** Data are *Missing at Random* if missingness depends on the values of the observed responses, but independent of the values of the missed responses.
**MCAR:** GLS estimator is still ML, whether based on all responses or on only “completers”, provided the mean is correctly specified;

**MAR:** Sample means and covariances based on either completers or all responses are biased, but ML is still valid provided means *and covariances* are correctly specified.